MOMENT CLOSURE IN A MORAN MODEL WITH RECOMBINATION

ELLEN BAAKE AND THIEMO HUSTEDT

Abstract. We extend the Moran model with single-crossover recombination to include general recombination and mutation. We show that, in the case without resampling, the expectations of products of marginal processes defined via partitions of sites form a closed hierarchy, which is exhaustively described by a finite system of differential equations. One thus has the exceptional situation of moment closure in a nonlinear system. Surprisingly, this property is lost when resampling (i.e., genetic drift) is included.

1. Introduction

In recent years, the processes of population genetics, which describe the genetic structure of populations under the influence of evolutionary forces such as mutation, selection, recombination, migration, and genetic drift, have been a rich source of fascinating probabilistic problems. More precisely, the dynamics is often well understood in the limit of infinite population size, where a law of large numbers leads to a deterministic description (in terms of discrete dynamical systems or differential equations), but great challenges ensue if the population is finite, in particular if there is interaction between individuals, such as competition (selection) or recombination (the combination of genetic material of two parents into the 'mixed' genetic type of an offspring); see [6, 9, 11]. Interactions usually make the infinite-population model nonlinear and, often, already difficult enough to treat. In the corresponding stochastic model, they are reflected by transition rates (or probabilities) that depend nonlinearly on the current state of the system and often result in processes whose treatment provides enormous challenges. Even the relationship between the stochastic process and its deterministic counterpart is usually unclear (apart from the infinite population limit). In particular, the expectation of the stochastic process is, usually, not given by the corresponding deterministic dynamics - in general, such coincidence is reserved for populations of individuals that evolve independently (as in branching processes); or systems with interactions that do not change the expectation (like Wright-Fisher sampling).

Indeed, even the analysis of the expectation is difficult in most processes of population genetics with interaction. Its dynamics does, usually, not only depend on the current expectation, but on higher moments, whose change, in turn, depends on even higher moments. Formulating this hierarchy of dependencies is a common approach for stochastic processes arising in various applications in physics, chemistry, and biology [15, 14, 8]. Usually, this hierarchy continues indefinitely (it does not 'close'); to extract at least an approximation to the (lower) moments of interest, some method of 'moment closure' must be employed (in the simplest case, a truncation) [8].

The corresponding deterministic systems (that arise through a law of large numbers) are also often tackled via systems of moments or cumulants, see [6, Ch. V.4] for an overview. Models of recombination take a special role between linear and nonlinear models. Although there is abundant interaction and hence nonlinearity, the deterministic system that describes the frequencies of all possible (geno)types may be (exactly) transformed into a linear one by embedding it into a higher-dimensional space (more explicitly, by adding further components that correspond to products of type frequencies). This method is known as Haldane linearisation [16]. The underlying linear structure even allows a diagonalisation and explicit solution, see [18] and references therein. In certain important special cases (notably, in so-called single-crossover dynamics in continuous time), this solution is surprisingly simple and immediately plausible [1, 3].

2000 *Mathematics Subject Classification.* Primary: 92D15; Secondary: 60J28.
Key words and phrases. Moment closure; Moran model; recombination.

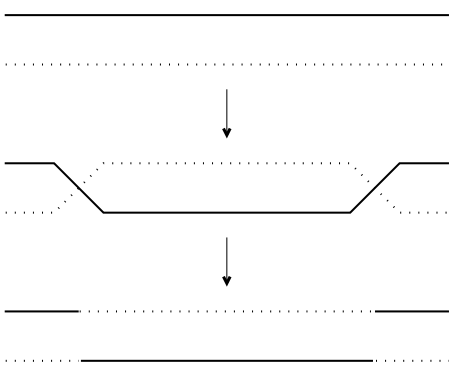

FIGURE 1. Recombination

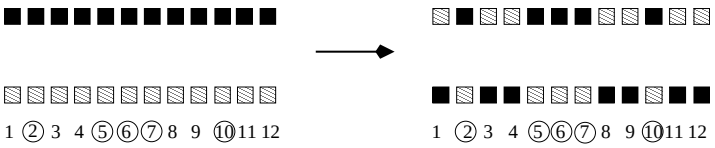

FIGURE 2. Recombination event defined by the set G (circled sites)

Elucidating underlying linear structures in the corresponding stochastic system (more precisely, in the Moran model with recombination) has only started very recently. In the aforementioned single-crossover case, Bobrowski et al. [5] analysed the asymptotic behaviour in the presence of mutation. Baake and Herms [4] observed that the expected type frequencies in the finite system (but without genetic drift) follow those in the deterministic model; this could be explained by the (conditional) independence of certain marginalised processes that appear as 'subsystems' of the stochastic model. This and other results now lead to the question whether in the general recombination scheme (i.e., not restricted to single crossovers) the dynamics of the expectations may be embedded into a higher but finite dimensional space, such that they are given by a finite system of differential equations? Is there an equivalent of Haldane linearisation in the sense of moments?

This article will address these questions in the framework of the Moran model with recombination and mutation. In particular it will show that the system of moments closes here after a finite number of steps, without any need for approximations, as long as there is no genetic drift. This may be considered as a stochastic analogue of Haldane linearisation.

2. Moran model with recombination

We consider a population of N individuals. Each of them is endowed with the set $S = \{1, \ldots, n\}$ of sites. These can be interpreted as nucleotide positions in a string of DNA or as gene loci on a chromosome. For each site i there is a finite set X_i of alleles that may occur at site i. A string of alleles is then called a type, $X := \bigtimes_{i=1}^{n} X_i$ is the type space.

We are interested in modelling recombination, which means the rearrangement of genetic material in sexually reproducing populations. It may occur during meiosis, the creation of gametes, that is egg cells or sperm. Homologous chromosomes may cross over at some points and exchange the genetic material in between (see Figure 1).

In the following we will assign recombination events to subsets of sites in a natural way. Let $G \subset S$. Then the corresponding recombination event between two individuals is the following: the alleles at the sites given by G remain at their positions, whereas the alleles at the sites in $\bar{G}$, the complement of G, are exchanged (see Figure 2).

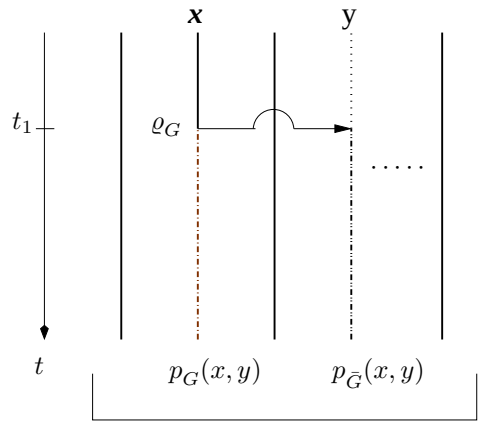

FIGURE 3. Moran model with recombination. At time t_1 the second individual, which is of type x, undergoes a recombination event corresponding to G and chooses its partner randomly, here the fourth individual, which is of type y; from that time on, the individuals are of type $p_G(x,y)$ and $p_{\bar{G}}(x,y)$.

We define the mappings $p_G : X \times X \to X$, $G \subset S$ by

$$p_G(x,y) =: \left(\bigtimes\nolimits_{i\in G}\{x_i\}\right) \times \left(\bigtimes\nolimits_{i\in \bar{G}}\{y_i\}\right) : , \tag{1}$$

where $: \cdots :$ means that the coordinates are ordered as in X. So, $p_G(x,y)$ and $p_{\bar{G}}(x,y)$ are the new types resulting from the recombination event corresponding to G between the types x and y. Obviously, $p_G(x,y) = p_{\bar{G}}(y,x)$. So, G and $\bar{G}$ essentially correspond to the same recombination event.

We now define a Moran model with recombination and mutation. Each individual undergoes recombination events corresponding to $G \subset S$ at rate $\varrho_G/4 \geq 0$ for all $G \subset S$. The recombination partner is chosen out of the whole population (including the opening individual itself). Then they exchange their genetic material according to the recombination event corresponding to G (see Figure 3). To keep things well-defined, the recombination rates ϱ_G have the properties $\varrho_G = \varrho_{\bar{G}}$ and $\varrho_\varnothing = \varrho_S = 0$.

Furthermore, mutation events may occur. An allele $x_i \in X_i$ at site i mutates into allele $y_i \in X_i$ with rate $\mu^i_{x_i y_i} \geq 0$. Thus, the mutation rate depends on both the parental and the offspring allele.

Additionally, we introduce birth events or, more precisely, resampling. Each individual produces an offspring at rate $b/2 \geq 0$. The offspring inherits the parent's type and replaces another individual, randomly chosen from the entire population (again including the parent individual).

In the following we are interested in the composition of the population, so we define the stochastic process $(Z_t)_{t\geq 0}$ with state space $E := \{\omega \text{ counting measure on } X \text{ with } \omega(X) = N\}$, by

$$Z_t(\{x\}) := \text{number of individuals of type } x.$$

In the following we will use shorthands like $Z_t(x)$, $z(x)$ instead of $Z_t(\{x\})$, $z(\{x\})$. Recombination, mutation and resampling events induce the following transitions if $Z_t = z$:

$$\begin{aligned} &z \to z + v_{G,x,y} \quad \text{with} \quad v_{G,x,y} := -\delta_x - \delta_y + \delta_{p_G(x,y)} + \delta_{p_{\bar{G}}(x,y)} \\ &\text{at rate} \quad \frac{1}{N}\varrho_G z(x) z(y) \quad \text{for} \quad x,y \in X, \quad G \subset S, \end{aligned} \tag{2}$$

$$z \to z - \delta_{(x_1,\ldots,x_i,\ldots,x_n)} + \delta_{(x_1,\ldots,y_i,\ldots,x_n)} \quad \text{at rate} \quad \mu^i_{x_i y_i} z(x), \tag{3}$$

$$z \to z + \delta_x - \delta_y \quad \text{at rate} \quad \frac{b}{2N} z(x)z(y). \tag{4}$$

The rate in (2) is determined in the following way: An individual of type x recombines at rate $\frac{1}{4}\varrho_G z(x)$ and chooses one individual of type y with probability $\frac{z(y)}{N}$. This leads to the rate $\frac{1}{4N}\varrho_G z(x)z(y)$ which needs to be multiplied by 4 to account for the fact that the recombination could be initiated by an individual of type y and that recombination according to G is the same as recombination according to $\bar{G}$.

A brief comment on the model is in order. We consider recombination and reproduction as independent events whereas, in true biology, recombination is coupled to reproduction. We use the decoupled version here because it is simpler, and because it allows to clearly separate the effects of random recombination from those of random reproduction. This version is also used elsewhere [17], on the argument that recombination events are rare.

For a subset $G \subset S$ we define $X_G := \bigtimes_{i\in G} X_i$ and the mapping $\pi_G : X \to X_G$ as the canonical projection. Let ω be a (signed) measure on X. We define the pullback $\pi_G.$ by $\pi_G.\omega := \omega \circ \pi_G^{-1}$. So, $\pi_G.$ maps a measure on X onto its corresponding marginal measure on X_G.

In the following, marginal processes of Z_t will play a crucial role. The following proposition states that these are Markov chains, too. It is an extension of Lemma 1 in [4].

Proposition 1. Let $I \subset S$, and let $(Z_t)_{t\geq 0}$ be the recombination process as defined by equations (2)-(4). Then $(\pi_I.Z_t)_{t\geq 0}$ is a Markov process with state space $E_I := \{\omega \text{ counting measure on } X_I, \omega(X_I) = N\}$.

Proof. Obviously, $(\pi_I.Z_t)_{t\geq 0}$ is a stochastic process on E_I.

We must show that the transition rates of $(\pi_I.Z_t)_{t\geq 0}$ only depend on the current state of the process. A recombination event induces the following transition:

$$\pi_I.z \to \pi_I.(z + v_{G,x,y}),$$

with

$$\pi_I.v_{G,x,y} = \delta_{\pi_I(p_G(x,y))} + \delta_{\pi_I(p_{\bar{G}}(x,y))} - \delta_{\pi_I(x)} - \delta_{\pi_I(y)} \tag{5}$$

and $\pi_I(p_G(x,y)) =: \Big(\bigtimes_{i\in G\cap I}\{x_i\}\Big) \times \Big(\bigtimes_{i\in \bar{G}\cap I}\{y_i\}\Big)$: in line with (1).

Consider now any nonzero jump. If it comes from a recombination event, it must be of the form (5). That means there are types $x_I, y_I \in X_I$ and a subset H of I such that $(\pi_I.z)(x_I)$ and $(\pi_I.z)(y_I)$ both decrease by one and the frequencies of the marginal types arising in the recombination event corresponding to H increase. The rate for this transition is then given by the sum of all transitions of the original process that induce this transition in the marginal process:

$$\begin{aligned}\sum_{\substack{G\subset S:\\ G\cap I=H}} \sum_{\substack{x\in X:\\ \pi_I(x)=x_I}} \sum_{\substack{y\in X:\\ \pi_I(y)=y_I}} \frac{\varrho_G}{N} z(x)z(y) &= \sum_{\substack{G\subset S:\\ G\cap I=H}} \frac{\varrho_G}{N} \big(\pi_I.z\big)\big(x_I\big)\cdot\big(\pi_I.z\big)\big(y_I\big) \\ &= \frac{\varrho_H^{(I)}}{N} \big(\pi_I.z\big)(x_I)\cdot\big(\pi_I.z\big)(y_I),\end{aligned} \tag{6}$$

with $\varrho_H^{(I)} := \sum_{G\subset S:\bar{G}\cap I=H} \varrho_G$. So, this last term depends only on the current state of the marginal process $(\pi_I.Z_t)_{t\geq 0}$.

A mutation event of an individual of type x at site i from allele x_i to allele y_i induces the following transition of $\pi_I.Z_t$:

$$\pi_I.z \to \pi_I.z + \pi_I.\delta_{(\dots,y_i,\dots)} - \pi_I.\delta_{(\dots,x_i,\dots)}.$$

This jump is zero if $i \notin I$. Obviously, the transition rate is $\mu^i_{x_i y_i}(\pi_I.z)(\pi_I(x))$ and depends merely on the current state of $\pi_I.Z_t$, too. The case of resampling is treated analogously. □

This proof is an example for the so-called lumping procedure for Markov chains, compare [7, 12] for the general context or [2] for the sequence context considered here.

Remark 1. A comparison between (2) and (6) shows that the marginal process $(\pi_I.Z_t)_{t\geq 0}$ can itself be considered as a recombination process on the sites I. So, assertions about Z_t will also hold for all derived marginal processes.

3. Recombination alone

In this Section we restrict ourselves to the case *without mutation and resampling*, that means with $\mu^i_{x_i y_i} = b = 0$ for all $i \in S$ and $x_i, y_i \in X_i$.

Since $v_{G,x,y} = 0$ for some $x, y \in X$, there are 'empty' recombination events at positive rate, but including these redundancies makes the rates in (2) so simple. The rates become considerably more complicated if only 'true jumps' are considered. This is already visible in the projection onto a single type. Let $x \in X$ and $Z_t = z$. In order to figure out the rate for the transition $z(x) \to z(x)+1$, we first determine the set of all pairs of types $\tilde{x}, \tilde{y} \in X$ such that for a given $G \subset S$ the jump $v_{G,\tilde{x},\tilde{y}}(x)$ equals 1:

$$\begin{aligned}\{\{\tilde{x},\tilde{y}\} \subset X : v_{G,\tilde{x},\tilde{y}}(x) = 1\} &= \{\{\tilde{x},\tilde{y}\} \subset X : \pi_G(\tilde{x}) = \pi_G(x), \pi_{\bar{G}}(\tilde{y}) = \pi_{\bar{G}}(x), \tilde{x} \neq x, \tilde{y} \neq x\} \\ &= \{\{\tilde{x},\tilde{y}\} \subset X : \tilde{x} \in \pi_G^{-1}\big(\pi_G(x)\big) \setminus \{x\}, \tilde{y} \in \pi_{\bar{G}}^{-1}\big(\pi_{\bar{G}}(x)\big) \setminus \{x\}\}.\end{aligned}$$

This leads to the transition rate

$$\sum_{G\subset S} \frac{\varrho_G}{2N}\big[\big(\pi_G.z\big)\big(\pi_G(x)\big) - z(x)\big]\big[\big(\pi_{\bar{G}}.z\big)\big(\pi_{\bar{G}}(x)\big) - z(x)\big]. \tag{7}$$

The transition rate for $z(x) \to z(x)-1$ can be figured out analogously; $v_{G,\tilde{x},\tilde{y}}(x) = -1$ iff $\tilde{x} = x$ and $\tilde{y}$ is any type which is neither in $\pi_G^{-1}\big(\pi_G(x)\big)$ nor in $\pi_{\bar{G}}^{-1}\big(\pi_{\bar{G}}(x)\big)$. Since $\pi_G^{-1}\big(\pi_G(x)\big) \cap \pi_{\bar{G}}^{-1}\big(\pi_{\bar{G}}(x)\big) = \{x\}$ one has the rate

$$\sum_{G\subset S} \frac{\varrho_G}{2N} z(x)\big[N - \big(\pi_G.z\big)\big(\pi_G(x)\big) - \big(\pi_{\bar{G}}.z\big)\big(\pi_{\bar{G}}(x)\big) + z(x)\big]. \tag{8}$$

Our aim is now to reformulate the process with the help of additional random variables, so that the transition rates become simpler, in particular, unaffected by empty events. To this end, we define two new counting measures derived from $(Z_t)_{t\geq 0}$, namely $(U_t)_{t\geq 0}$ by $U_0(x) = 0$ and

$$U_t(x) = \text{number of events at which } x\text{-individuals are created until time } t$$

and $(V_t)_{t\geq 0}$ by $V_0(x) = 0$ and

$$V_t(x) = \text{number of events at which } x\text{-individuals are broken up until time } t.$$

These processes also count events at which Z_t does not change, namely the case that individuals of type x are created and broken up at the same time. This may happen when an individual of type x recombines according to G with an individual of type y with $\pi_G(y) = \pi_G(x)$. Whenever this occurs, both counters increase but their difference remains unchanged. Altogether, we thus have

$$Z_t = Z_0 + U_t - V_t \tag{9}$$

with the transition rates of U_t and V_t unaffected by 'empty' events: For $U_t(x) = u$, $u \to u+1$ happens at rate

$$\sum_{G\subset S} \frac{\varrho_G}{2N}\big(\pi_G.z\big)\big(\pi_G(x)\big) \cdot \big(\pi_{\bar{G}}.z\big)\big(\pi_{\bar{G}}(x)\big),$$

and for $V_t(x) = v$, the transition $v \to v+1$ happens at rate

$$\sum_{G\subset S} \frac{\varrho_G}{2} z(x).$$

In the following, marginal processes will emerge frequently. We introduce a short-hand, symbolic notation similar to the one described in [1]. Fix an arbitrary $x \in X$ and define for a subset $G = \{g_1, \ldots, g_{|G|}\}$ of sites

$$[G]_t := [g_1, \ldots, g_{|G|}]_t := \big(\pi_G.Z_t\big)\big(\pi_G(x)\big).$$

$[G]_t$ is the number of individuals that are identical to x at the sites corresponding to G, at time t. Again, we use shorthands $[g_1, \ldots, g_{|G|}]_t$ instead of $[\{g_1, \ldots, g_{|G|}\}]_t$. Note that we suppress the dependence on x in $[G]_t$ for ease of notation. Analogously, we define for the processes U_t and V_t:

$$\langle G\rangle_t := \big(\pi_G.U_t\big)\big(\pi_G(x)\big),$$

$$(G)_t := \big(\pi_G.V_t\big)\big(\pi_G(x)\big).$$

By Remark 1, we can now consider $[G]_t$ as a recombination process on $|G|$ sites evaluated at the type $(x_{g_1}, \ldots, x_{g_{|G|}})$.

For $|G| = 2$, the distribution of $\langle g_1, g_2\rangle_t$ can be given explicitly because the transition rates

$$\sum_{H\subset S:|H\cap G|=1} \frac{\varrho_H}{2N}[g_1]_t[g_2]_t =: a \tag{10}$$

are constant in time because all 1-site marginals are constant in time. So, $\langle g_1, g_2\rangle_t$ follows a Poisson distribution with parameter at.

3.1. **Analysis of the expectation.** Since we will use it frequently, we want to recall an elementary fact concerning the dynamics of the mean of a continuous-time Markov chain with a finite state space, which is often used implicitly. The proof is a straightforward exercise that can be found in [4, Fact 1], for example.

Lemma 1. Let $(Z_t)_{t\geq 0}$ be a Markov process with finite state space $E \subset \mathbb{Z}^d$ with transition rates $q(z, z+v)$ for transitions from z to $z+v$ for $z \in E$, $v \neq 0$ (let $q(z, z+v) = 0$ if $z+v \notin E$). Then the following equation holds for all $t \geq 0$

$$\frac{d}{dt}\mathbb{E}(Z_t) = \mathbb{E}(F(Z_t)),$$

where F is defined as

$$F(z) := \sum_{v\in\mathbb{Z}^d} vq(z, z+v).$$

□

Lemma 1 together with the representation of Z_t in (9) gives us the dynamics of the mean:

$$\frac{d}{dt}\mathbb{E}\Big[[1, \ldots, n]_t\Big] = \sum_{G\subset S} \mathbb{E}\Big[\frac{\varrho_G}{2N}([G]_t[\bar{G}]_t - N \cdot [1, \ldots, n]_t)\Big]. \tag{11}$$

The motivation for this comes from the well-understood special case of single crossovers [4]. Here, all recombination rates that are attached to multiple crossover recombination events vanish. This affects all ϱ_G with G that either do not contain 1 or n, or have gaps.

In this case, the induced marginal processes are conditionally independent of each other and so moment closure is immediate [4, Lemma 1 and Theorem 1]:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\Big[[1, \ldots, n]_t\Big] &= \sum_{G\subset S} \mathbb{E}\Big[\frac{\varrho_G}{2N}([G]_t[\bar{G}]_t - N \cdot [1, \ldots, n]_t)\Big] \\ &= \sum_{G\subset S} \frac{\varrho_G}{2N}(\mathbb{E}\Big[[G]_t\Big]\mathbb{E}\Big[[\bar{G}]_t\Big] - N \cdot \mathbb{E}\Big[[1, \ldots, n]_t\Big]).\end{aligned}$$

We obtain a finite nonlinear system of differential equations, whose solution is known in closed form [1].

The independence relies on two properties. First, a single crossover recombination event induces a pair of marginal processes $[G]_t, [\bar{G}]_t$ for which $\{G, \bar{G}\}$ is an *ordered* partition of S. Second, a single crossover recombination event only affects one of the induced processes while leaving the other one constant.

With general recombination both these properties are violated. First, marginal processes arise that are given by *non-ordered* partitions, so even single-crossover recombination events may affect both processes at the same instant. Second, a multiple-crossover recombination event may affect

the frequency of a pair of marginals that are given by an ordered partition. So, the independence of the induced marginal processes is violated in two ways.

Let us now look at (11) again. On the right-hand side, an expectation of products emerges. This is what one may expect due to the inherent nonlinearity of the recombination process. Nevertheless, we see that no site arises more than once, so the arising products are described by a partition of sites. This leads us to the following question: Given an arbitrary partition of sites, what is the dynamics of the mean of the product of the induced marginal processes? Theorem 1 below answers this. For its formulation we need the following definition.

Definition 1. Let $\{A_j\}_{j\in J}$ be a collection of sets with $A_i \cap A_j = \varnothing,\ i \neq j$. Define $A_J := \bigcup_{j\in J} A_j$. Then, $G \subset A_J$ *disrupts* $\{A_j\}_{j\in J}$, denoted by $G|\{A_j\}_{j\in J}$, if $G \cap A_j \notin \{\varnothing, G\}$ for all $j \in J$. For $|J| = 1$, we simply write $G|A_j$.

Note that for a collection of pairwise disjoint subsets of sites $\{A_j\}_{j\in J}$ disrupted by G, in a recombination event corresponding to G between individuals of marginal types $\pi_G(x)$ and $\pi_{A_J\setminus G}(x)$, the processes $\langle A_j\rangle_t,\ j \in J$ increase. Similarly, in the recombination event corresponding to G between individuals of marginal types $\pi_{A_K}(x)$ and $\pi_{A_J\setminus A_K}(x)$ for $K \subset J$, the processes $(A_j)_t,\ j \in J$ increase.

With these preparations, we are now ready to state

Theorem 1. *Let $m \leq n$, $M := \{1,\dots,m\}$ and $\mathcal{A} := \{A_1,\dots,A_m\}$ be a partition of $\{1,\dots,n\}$. Define $\mathcal{P}$ as the set of all triples (I,J,K), where $\{I,J,K\}$ is a partition of M. Then*

$$\frac{d}{dt}\mathbb{E}\Big[\prod_{\ell\in M}[A_\ell]_t\Big] = \mathbb{E}\Big[\sum_{\substack{(I,J,K)\in\mathcal{P}\\ I\neq M}}\prod_{i\in I}[A_i]_t \sum_{\tilde{K}\subset K}\sum_{\substack{G\subset A_J\\ G|\{A_j\}_{j\in J}}} \frac{\varrho^I_{K,G}}{4N}(-1)^{|K|}[A_{\tilde{K}}\cup G]_t[A_{K\setminus\tilde{K}}\cup G^c]_t\Big], \tag{12}$$

where G^c is the complement of G in A_J and $\varrho^I_{K,G}$ is defined as

$$\varrho^I_{K,G} := \sum_{D\subset A_I}\sum_{\substack{H\subset A_K\\ H|\{A_k\}_{k\in K}}} \varrho_{H\cup D\cup G}. \tag{13}$$

Remark 2. The right-hand side of (12) may be read in the following way. The set I indicates the parts of $\mathcal{A}$ that remain unchanged under the corresponding recombination event, the sets J and K indicate sets for which the derived processes U_t and V_t, respectively, increase. So the splitting of Z_t into U_t and V_t does not only simplify the calculation but also shows up in the result.

Proof of Theorem 1. For $\delta t > 0$, define

$$\langle A_\ell\rangle^t_{\delta t} := \langle A_\ell\rangle_{t+\delta t} - \langle A_\ell\rangle_t$$

and

$$(A_\ell)^t_{\delta t} := (A_\ell)_{t+\delta t} - (A_\ell)_t$$

then

$$[A_\ell]_{t+\delta t} = [A_\ell]_t + \langle A_\ell\rangle^t_{\delta t} - (A_\ell)^t_{\delta t}$$

and $\prod_{\ell\in M}[A_\ell]_{t+\delta t}$ reads

$$\prod_{\ell\in M}[A_\ell]_{t+\delta t} = \sum_{(I,J,K)\in\mathcal{P}}(-1)^{|K|}\prod_{i\in I}[A_i]_t\prod_{j\in J}\langle A_j\rangle^t_{\delta t}\prod_{k\in K}(A_k)^t_{\delta t}.$$

Let $t+\delta t$ be the time of the first recombination event after time t.
Then, a summand $\prod_{i\in I}[A_i]_t\prod_{j\in J}\langle A_j\rangle^t_{\delta t}\prod_{k\in K}(A_k)^t_{\delta t}$ may evaluate to:

- zero if there is any $j \in J$ or $k \in K$ such that $\langle A_j\rangle^t_{\delta t} = 0$ or $(A_k)^t_{\delta t} = 0$
- $(-1)^{|K|}\prod_{i\in I}[A_i]_t$ otherwise, that means if $\langle A_j\rangle^t_{\delta t} = (A_k)^t_{\delta t} = 1$ for all $j \in J$, $k \in K$.

The latter transition comes from recombination events that correspond to the union of some G disrupting $\{A_j\}_{j\in J}$ and H disrupting $\{A_k\}_{k\in K}$ and any subset D of A_I. At such recombination events, the recombining individuals must be of the following form: x-alleles at G, G^c resp., x-alleles

at $A_k,\ k \in K$, whereas the particular A_k may be arbitrarily distributed across the two individuals (but the individual sets may not be disrupted!). Thus, the complete rate reads

$$r(I,J,K) := \sum_{\tilde{K}\subset K} \sum_{\substack{G\subset A_J \\ G|\{A_j\}_{j\in J}}} \frac{\varrho^I_{K,G}}{4N} [A_{\tilde{K}} \cup G]_t [A_{K\setminus\tilde{K}} \cup G^c]_t, \tag{14}$$

with G^c and $\varrho^I_{K,G}$ as defined above. This is the rate of the event that the terms corresponding to J and K increase, that means it is the rate of all recombination events such that a binding arises in each $A_j,\ j \in J$, and a binding breaks in each $A_k,\ k \in K$.

Thus,

$$\frac{d}{dt}\mathbb{E}\Big[\prod_{\ell\in M}[A_\ell]_t\Big] = \mathbb{E}\Big[\sum_{\substack{(I,J,K)\in\mathcal{P} \\ I\neq M}} (-1)^{|K|}\prod_{i\in I}[A_i]_t\, r(I,J,K)\Big],$$

which is the assertion of the theorem.

□

Let us now consider the implication of the theorem for the moment-closure problem. The theorem tells us that the dynamics of the mean of a product of marginal processes defined by a partition of sites can be described by the mean of another product of marginal processes defined by a (finer) partition of sites. Since the number of sites is finite and so is the number of partitions of sites the moment closure approach (for the mean) directly leads to a finite and linear system of ODE's. We have thus proved

Corollary 1. For the Moran model with recombination alone, the moment approach closes. □

The size of these systems explodes with the number of sites. Nevertheless, there is much redundancy in the concrete calculation of particular means. For example, in the analysis of $\mathbb{E}\big[[1,2,3,4]_t\big]$, marginal processes on three sites emerge. According to Remark 1, these can be treated as recombination processes on three sites, so by a proper summation of the recombination rates, one can easily determine their solutions given the solution of the three-sites recombination process.

3.2. **Comparison with the deterministic dynamics.** We now want to compare the result of Theorem 1 to the corresponding deterministic dynamics. To this end, let $\mathcal{M}(X)$ be the space of all measures on X. For $G \subset S$ define the recombinator R_G [1] by

$$R_G(\omega) := \frac{1}{|\omega|}(\pi_G.\omega) \otimes (\pi_{\bar{G}}.\omega)$$

with $R_G(0) = 0$. Consider the following dynamical system on $\mathcal{M}(X)$:

$$\dot{\omega} = \sum_{G\subset S} \frac{\varrho_G}{2}\big(R_G - \mathbb{1}\big)\omega. \tag{15}$$

This is the infinite population limit of the recombination process (without and with resampling) in the following sense. If we consider $\hat{Z}^N_t := \frac{1}{N}Z_t$ and let $\lim_{N\to\infty} Z^N_0 = p_0$, then

$$\lim_{N\to\infty} \sup_{s\le t} |\hat{Z}^N_s - p_s| = 0 \tag{16}$$

with probability 1, where p_s is the solution of the initial value problem (15) with $\omega_0 = p_0$. This is shown in [4] for the special case of single crossovers, but it is obvious that the proof, which is based on the general law of large numbers by Ethier and Kurtz ([10, Thm. 11.2.1], see also [13]), may be generalised to the case of multiple crossovers.

[1] This is a generalisation of the recombinator in [1]. Note that the notational similarity is deceptive because G denotes sites here rather than 'links' (the bonds between sites) as in [1].

We are now interested in the relationship between (12) and the deterministic dynamics. If ω_t follows (15), then a (tensor) product of marginal measures $(\pi_{A_1}.\omega_t)\otimes\cdots\otimes(\pi_{A_m}.\omega_t)$ given by a partition of sites as in Theorem 1 exhibits the following dynamics:

(17)
$$\begin{aligned}\frac{d}{dt}\big((\pi_{A_1}.\omega_t)\otimes\cdots\otimes(\pi_{A_m}.\omega_t)\big) &= (\pi_{A_1}.(\sum_{G\subset S}\frac{\varrho_G}{2}(R_G-\mathbb{1}))(\omega_t))\otimes(\pi_{A_2}.\omega_t)\otimes\cdots\otimes(\pi_{A_m}.\omega_t)\\ &+(\pi_{A_1}.\omega_t)\otimes(\pi_{A_2}.(\sum_{G\subset S}\frac{\varrho_G}{2}(R_G-\mathbb{1}))\omega_t)\otimes(\pi_{A_3}.\omega_t)\otimes\cdots\otimes(\pi_{A_m}.\omega_t)\\ &+\cdots+(\pi_{A_1}.\omega_t)\otimes\cdots\otimes(\pi_{A_{m-1}}.\omega_t)\otimes(\pi_{A_m}.(\sum_{G\subset S}\frac{\varrho_G}{2}(R_G-\mathbb{1}))\omega_t)\\ &=\sum_{j=1}^{m}\sum_{B\subset A_j}\varrho_B(\pi_{A_1}.\omega_t)\otimes\cdots\otimes\big[\frac{1}{|\omega_t|}(\pi_B.\omega_t)\otimes(\pi_{A_j\setminus B}.\omega_t)-(\pi_{A_j}.\omega_t)\big]\otimes\cdots\otimes(\pi_{A_m}.\omega_t),\end{aligned}$$

with $\varrho_B := \sum_{\substack{H\subset S\\ H\cap A_j=B}}\varrho_H$.

Compare this to (12), and only consider summands where $|J|=1$ and $|K|=0$ or $|J|=0$ and $|K|=1$. According to Remark 2, we can understand the corresponding transitions as 'uncorrelated' events at which only one marginal process changes at a given instant. We get the following terms on the right-hand side of (12):

- $J=\{j\}, K=\varnothing$:

$$\mathbb{E}\Big[\prod_{i:i\neq j}[A_i]_t\sum_{\substack{G\subset A_j\\ G|A_j}}\frac{\varrho^{M\setminus\{j\}}_{\varnothing,G}}{4N}[G]_t[A_j\setminus G]_t\Big], \tag{18}$$

- $J=\varnothing, K=\{j\}$:

$$\mathbb{E}\Big[\prod_{i:i\neq j}[A_i]_t\frac{\varrho^{M\setminus\{j\}}_{\{j\},\varnothing}}{4N}(-1)[A_j]_tN\Big], \tag{19}$$

with (cf. (13))

$$\varrho^{M\setminus\{j\}}_{\{j\},\varnothing}=\sum_{D\subset A_I}\sum_{\substack{H\subset A_j\\ H|A_j}}\varrho_{H\cup D}=\sum_{\substack{H\subset A_j\\ H|A_j}}\sum_{D\subset A_I}\varrho_{H\cup D}=\sum_{\substack{H\subset A_j\\ H|A_j}}\varrho^{M\setminus\{j\}}_{\varnothing,H}.$$

Adding all terms of kind (18) and (19), one obtains the analogue of the right-hand side of (17). According to Remark 2, the summands with $|J\cup K|\geq 2$ correspond to 'correlated' events at which two or more marginal processes change simultaneously.

We may thus conclude that the uncorrelated events correspond to the deterministic equation. We will now show that the correlated events are of lower order and thus tend to zero in the limit $N\to\infty$. To this end, look at (11). The right-hand side consists of the terms $\frac{1}{N}[G]_t[\bar{G}]_t$ and $[1,\dots,n]_t$. They are both of order N, since each individual term $[\dots]_t$ is of order N (which follows, for example, from (16)). Let us look at the derivative of the mean of $\frac{1}{N}[G]_t[\bar{G}]_t$ (cf. (12)). The terms with $|I|=1$ (those belonging to 'uncorrelated' events) will be of order N again, whereas the terms with $I=\varnothing$ are of order 1. By differentiating terms such as $\mathbb{E}\Big[\frac{1}{N^2}[A_1]_t[A_2]_t[A_3]_t\Big]$ and beyond, the same observation applies: the order of summands belonging to 'correlated' events is less or equal 1, so for the relative frequencies $([1,\dots,n]_t/N)$ the dynamics of the mean tends to the dynamics of the deterministic model.

3.3. **Two sites, arbitrary moments.** In the case of two sites, the recombination process is rather simple. This mainly relies on the fact that the transition rate for $\langle 1,2\rangle_t$ is constant, as we have already seen in (10). Furthermore the set of partitions of two sites is trivial. In this special case we can easily show moment closure for arbitrary moments. The simplicity of the setting

permits to look at $[1,2]_t$ itself without considering $\langle 1,2\rangle_t$ and $(1,2)_t$. The process $[1,2]_t^m$ has the following possible transitions (cf. (7), (8)):

$$[1,2]_t^m \to ([1,2]_t+1)^m \quad \text{at rate} \quad \frac{\varrho_1}{N}([1]_t-[1,2]_t)([2]_t-[1,2]_t)$$

and

$$[1,2]_t^m \to ([1,2]_t-1)^m \quad \text{at rate} \quad \frac{\varrho_1}{N}[1,2]_t(N-[1]_t-[2]_t+[1,2]_t).$$

Using the binomial theorem and eliminating empty transitions, we obtain for the m-th moment:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\Big[[1,2]_t^m\Big] =& \frac{\varrho_1}{N}\sum_{k=0}^{m-1}\mathbb{E}\Big[\binom{m}{k}[1,2]_t^k([1]_t-[1,2]_t)([2]_t-[1,2]_t)\Big]\\ &+\frac{\varrho_1}{N}\sum_{k=0}^{m-1}(-1)^{m-k}\mathbb{E}\Big[\binom{m}{k}[1,2]_t^k[1,2]_t(N-[1]_t-[2]_t+[1,2]_t)\Big]\\ =&\sum_{k=0}^{m-2}\mathbb{E}\Big[\frac{\varrho_1}{N}\binom{m}{k}[1,2]_t^k\{[1]_t[2]_t-2\delta^{(2)}_{m-k}[1,2]_t c+2\delta^{(2)}_{m-k+1}[1,2]_t^2+(-1)^{m-k}[1,2]_t N\}\Big]\\ &+\frac{\varrho_1 m}{N}\mathbb{E}\Big[[1,2]^{m-1}([1]_t[2]_t-[1,2]_t N)\Big],\end{aligned}$$

with $c:=([1]_t+[2]_t)$ and

$$\delta^{(2)}_{m-k}:=\begin{cases}1 & \text{if} \quad m-k\equiv 0 \bmod 2\\ 0 & \text{otherwise.}\end{cases}$$

So all emerging terms are moments of order m or less.

4. Recombination and Mutation

We now want to add mutation to our process. Let us first look at the process with mutation alone, e.g. $b=\varrho_G=0$. By Lemma 1, the derivative of the mean is:

$$\frac{d}{dt}\mathbb{E}\Big[[1,\dots,n]_t\Big]=\sum_{j\in S}\Big(\sum_{y\in X_j}\mu^j_{yx_j}[1,\dots,j-1,j+1,\dots,n]_t-\sum_{y\in X_j\setminus\{x_j\}}\mu^j_{x_jy}[1,\dots,n]_t\Big).$$

So, it only consists of linear terms and marginal processes. When we consider a product of marginal processes given by a partition of sites as in the previous section, we have, due to the fact that mutation only acts on single sites independently of others:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\Big[\prod_{\ell\in M}[A_\ell]_t\Big]=\mathbb{E}\Big[&\sum_{\ell\in M}\Big(\prod_{i\in M\setminus\{\ell\}}[A_i]_t\sum_{j\in A_\ell}\sum_{y\in X_j}\mu^j_{yx_j}[A_\ell\setminus\{j\}]_t\Big)\\ &-\sum_{\ell\in M}\Big(\prod_{i\in M\setminus\{\ell\}}[A_i]_t\sum_{j\in A_\ell}\sum_{y\in X_j\setminus\{x_j\}}\mu^j_{x_jy}[A_\ell]_t\Big)\Big].\end{aligned}$$

What happens when we add recombination? Let F_M and F_R, respectively, be the 'mean rate of change functions' from Lemma 1 for $\prod_{\ell\in M}[A_\ell]_t$ from the process with solely mutation and recombination, respectively. Since mutation and recombination proceed independently, the respective function F_{RM} for the recombination-mutation process is then just F_R+F_M and according to Lemma 1 we have:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\Big[\prod_{\ell\in M}[A_\ell]_t\Big]=\mathbb{E}\Big[&\sum_{\substack{(I,J,K)\in\mathcal{P}\\ I\neq M}}\prod_{i\in I}[A_i]_t(-1)^{|K|}\sum_{\tilde K\subset K}\sum_{\substack{G\subset A_J\\ G|\{A_j\}_{j\in J}}}\frac{\varrho^I_{K,G}}{4N}[A_{\tilde K}\cup G]_t[A_{K\setminus\tilde K}\cup G^c]_t\\ &+\sum_{\ell\in M}\Big(\prod_{i\in M\setminus\{\ell\}}[A_i]_t\sum_{j\in A_\ell}\sum_{y\in X_j}\mu^j_{yx_j}[A_\ell\setminus\{j\}]_t\Big)\\ &-\sum_{\ell\in M}\Big(\prod_{i\in M\setminus\{\ell\}}[A_i]_t\sum_{j\in A_\ell}\sum_{y\in X_j\setminus\{x_j\}}\mu^j_{x_jy}[A_\ell]_t\Big)\Big].\end{aligned}$$

So, the arising terms are the same as in the pure recombination process plus linear terms. It is therefore clear that we have moment closure here as well.

5. Recombination and Resampling

In this section, we set $b > 0$ and the mutation rates zero again, so we look at the Moran model with recombination and resampling only. At first glance, one may think that resampling has no effect on the expectation, since the process with resampling alone has a constant mean. Indeed, the first derivative of the mean looks the same as in the pure recombination case:

$$\frac{d}{dt}\mathbb{E}\big[[1,2]_t\big] = \frac{\varrho_1}{N}\mathbb{E}\big[[1]_t[2]_t - [1,2]_t N\big]. \tag{20}$$

However, due to resampling, the one-site marginal processes are no longer constant, so we do not have instantaneous moment closure any more (cf. (4): at a resampling event, the frequency of alleles may change). The derivative of their product is obtained after an elementary but lengthy calculation:

$$\frac{d}{dt}\mathbb{E}\big[[1]_t[2]_t\big] = \frac{b}{N}\mathbb{E}\big[[1,2]_t N - [1]_t[2]_t\big]. \tag{21}$$

We obtain a finite linear system of differential equations, namely (20) and (21). In particular, we obtain

$$\frac{d}{dt}\mathbb{E}\big[[1,2]_t N - [1]_t[2]_t\big] = -\big(\frac{\varrho_1}{N} + \frac{b}{N}\big)\mathbb{E}\big[[1,2]_t N - [1]_t[2]_t\big] \tag{22}$$

with the obvious exponential solution. The term $[1,2]_t N - [1]_t[2]_t$ is a correlation function, a so called linkage disequilibrium, which is widely used in population genetics. We see that both, recombination and resampling, reduce correlations between sites.

For more than two sites, exact moment closure can no longer be established. To make this plausible, we will only present the derivative of $\mathbb{E}\big[[1]_t[2]_t[3]_t\big]$ (again, the calculation is elementary but lengthy), which is a term that emerges in the derivatives of the process on three sites due to recombination:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\big[[1]_t[2]_t[3]_t\big] = \frac{b}{N}\mathbb{E}\big[&[1]_t[2,3]_t N + [2]_t[1,3]_t N + [3]_t[1,2]_t N + [1]_t[1,2]_t[1,3]_t \\ &+ [2]_t[1,2]_t[2,3]_t + [3]_t[1,3]_t[2,3]_t - 3[1]_t[2]_t[3]_t - [1]_t^2[1,2,3]_t - [2]_t^2[1,2,3]_t - [3]_t^2[1,2,3]_t\big].\end{aligned}$$

The last three terms are quadratic and it is clear that further differentiating will lead to terms such as $[1]_t^3[2]_t[3]_t$ whose derivative will contain moments of even higher order. Thus, the interaction between recombination and resampling destroys moment closure.

6. Conclusion

In this paper, we have extended the single-crossover Moran model from [4] to include general recombination. The dynamics of the expectation under general recombination becomes significantly more complicated. In particular, it now deviates from the dynamics in the infinite population model. The reason is the loss of independence of certain marginal processes.

As is usual with nonlinear processes, the dynamics of a given moment requires higher moments. Nevertheless, in this case after a finite number of steps no additional terms emerge. This is due to the fact that the arising processes may in each step be described by a partition of sites. When mutation is included, this exact moment closure persists, but the arising processes can no longer be described by a partition of sites. Altogether, we have an exception to the rule that the dynamics of the moments of nonlinear processes lead to infinite hierarchies of ODE's.

This exact moment closure gets lost when we extend the model to include genetic drift (i.e., resampling). This is, of course, disappointing since the Moran model with recombination alone is mathematically interesting, but of limited biological value. Nevertheless, the resulting hierarchy of moments might be interesting to analyse with respect to the various possibilities of approximate moment closure.

Furthermore, the arising terms such as $\mathbb{E}\big[\prod_{\ell\in M}[A_\ell]_t\big]$ are of considerable interest in population genetics beyond this moment closure procedure, since they are the building blocks of the linkage disequilibria [6] that are so important in population genetics (compare (22) for the simplest example).

References

[1] Baake, M., Baake, E. (2003). An exactly solved model for mutation, recombination and selection. *Can. J. Math.* **55** 3–41 and **60** (2008), 264–265 (Erratum); `arXiv:math.CA/0210422`.

[2] Baake, E., Baake, M., Bovier, A., Klein, M. (2005). An asymptotic maximum principle for essentially nonlinear evolution models. *J. Math. Biol.* **50,** 83–114; `arXiv:q-bio/0311020v2`.

[3] Baake, M. (2005). Recombination semigroups on measure spaces. *Monatsh. Math.* **146,** 267–278 and **150** (2007), 83–84 (Addendum); `arXiv:math.CA/0506099`.

[4] Baake, E., Herms, I. (2008). Single-crossover dynamics: finite versus infinite populations, *Bull. Math. Biol.* **70,** 603–624; `arXiv:q-bio/0612024v2`.

[5] Bobrowski, A., Wojdyla, T., Kimmel, M. (2010). Asymptotic behavior of a Moran model with mutations, drift and recombination among multiple loci. *J. Math. Biol.* **61,** 455–473.

[6] Bürger, R. (2000). *The Mathematical Theory of Selection, Recombination and Mutation.* Wiley, Chichester.

[7] Burke, C., Rosenblatt, M. (1958). A Markovian function of a Markov chain. *Ann. Math. Stat.* **29,** 1112–1122.

[8] Dieckmann, U., Law, R. (2000). Relaxation projections and the method of moments. In: The Geometry of Ecological Interactions: Simplifying Spatial Complexity, eds Dieckmann, U., Law, R., Metz, J.A.J., Cambridge University Press, 412–455.

[9] Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*, 2nd edn. Springer, New York.

[10] Ethier, S. N., Kurtz, T. G. (1986) *Markov Processes - Characterization and Convergence*, Wiley, New York. Reprint 2005.

[11] Ewens, W. (2004). *Mathematical Population Genetics*, 2nd edn. Springer, Berlin.

[12] Kemeney, J.G., Snell, J.L. (1981). *Finite Markov Chains.* Springer, New York.

[13] Kurtz, T. G. Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes. *J. Appl. Prob.* **8,** 344–356.

[14] Lee, C.H., Kim, K. and Kim, P. (2009). A moment closure method for stochastic reaction networks. *J. Chem. Phys.* **130,** 134107.

[15] Levermore, C.D. (1996). Moment closure hierarchies for kinetic theories. *J. Stat. Phys.* **83,** 1021–1065.

[16] McHale, D., Ringwood, A. (1983). Haldane linearisation of baric algebras. *J. London Math. Soc.* (2) **28,** 17–26.

[17] Pfaffelhuber, P., Haubold, B., Wakolbinger, A. (2006). Approximate genealogies under genetic hitchhiking. *Genetics* **174,** 1995–2008.

[18] von Wangenheim, U., Baake, E., Baake, M. (2010). Single–crossover recombination in discrete time. *J. Math. Biol.* **60,** 727–760; `arXiv:0906.1678v1`.

(Ellen Baake) Technische Fakultät, Universität Bielefeld, Box 100131, 33501 Bielefeld, Germany
E-mail address, Ellen Baake: `ebaake@techfak.uni-bielefeld.de`

(Thiemo Hustedt) Forschungsschwerpunkt Mathematisierung, Universität Bielefeld, Box 100131, 33501 Bielefeld, Germany
E-mail address, Thiemo Hustedt: `thustedt@uni-bielefeld.de`